A two-stage sampling strategy improves chamber-based estimates of greenhouse gas fluxes

He, Yufeng; Gibbons, James; Rayment, Mark

Agricultural and Forest Meteorology

DOI:
10.1016/j.agrformet.2016.06.015

Published: 15/11/2016

Peer reviewed version

A two-stage sampling strategy improves chamber-based estimates of greenhouse gas fluxes

Yufeng He[1], James Gibbons[1], and Mark Rayment[1]

[1]School of Environment, Natural Resources and Geography, Bangor University, Bangor, UK.
Corresponding author: Yufeng He (afp23e@bangor.ac.uk)

Note:

This is a **post-print version** of the article.

The published version:

Yufeng He, James Gibbons, Mark Rayment, A two-stage sampling strategy improves chamber-based estimates of greenhouse gas fluxes, Agricultural and Forest Meteorology, Volumes 228–229, 15 November 2016, Pages 52-59, ISSN 0168-1923, http://dx.doi.org/10.1016/j.agrformet.2016.06.015.

Available online: http://www.sciencedirect.com/science/article/pii/S0168192316303124

Abstract

Fluxes of greenhouse gases (GHG) are typically characterized by high spatial and temporal variability and large sample sizes (e.g. >30) are thus required to obtain a reliable estimate of the population mean and variance when using simple random sampling (SRS). Sample size, however, is often constrained by budget (time, labor) and therefore practical considerations induce significant (but unknown) measurement error and bias from sampling. In this paper we report a two-stage sampling strategy (2SS) by which the same level of sampling accuracy achievable by SRS can be achieved with significantly smaller sample sizes by optimizing sub-sample selection to retain the statistical characteristics of the sample population. Comparisons between 2SS and SRS were conducted using three datasets with low, medium and high coefficients of variance (*CV*). The size of the first (n') and second (n) stage samples had significant effects on overall sample accuracy. Across all datasets, 2SS reduced RMSE of mean and variance by an average of 30%. The absolute reduction in RMSE of mean and variance was found to be nearly proportional to the value of *CV*, such that the dataset with the largest *CV* showed the largest benefit from 2SS. Logarithmic relationships were found between the difference in the RMSEs and the ratio, n'/ n, serving as a guide to allocate sampling resources in practice. Employing 2SS will aid accurate quantification of soil GHG fluxes in all but the most homogeneous situations.

Keywords: GHG flux, sampling, chamber-based measurement, spatial variability

1 Introduction

Chamber-based measurements of the flux of greenhouse gases (GHG) emissions from soils at local scales (less than 1 km^2) are a pillar of Kyoto reporting, especially in agriculture and

land use, land-use change and forestry (IPCC 2000). Emission factors generated from chamber-based measurements of total or mean GHG emissions from land use categories are typically based on relatively few measurements in time and space. Errors or uncertainty in the quantified emissions are directly and linearly propagated into the total national accounts. What confidence do we have in the accuracy of our estimates? Very little, especially for CH_4 and N_2O where uncertainty spans orders of magnitude (Maljanen et al. 2010; Rayment & Jarvis 2000; Rochette & Eriksen-Hamel 2008; Venterea et al. 2009).

Comparisons of chamber measurements, scaled to the field scale, with eddy covariance (EC) measurements directly measuring at the field scale (i.e. two methods purporting to measure the same thing) often reveal large and unsystematic differences (Davidson *et al*., 2002; Goulden *et al*., 1996; Jones *et al*., 2011; Reth *et al*., 2005). However estimates of, for example, annual net fluxes are typically presented with uncertainty bounds so large as to suggest that the estimates are, in fact, in agreement. Without suggesting that either chamber-based measurements or EC-based estimates are inherently better than the other, it is arguable that the EC community have confronted measurement uncertainty squarely and openly (Baldocchi 2003; Hollinger & Richardson 2005; Oren et al. 2006), and have produced methodologies for assessing and reporting uncertainties, directed towards the ultimate aim of reducing them (Baldocchi et al. 2000; Gu et al. 2012; Foken et al. 2004). On the other hand, the chamber-based measurement community, though revealing error sources from decades of experience has been slower to explore measurement uncertainty caused by sampling (Davidson et al. 2002).

Amongst the literature there are many attempts to grapple with the surrounding chamber design and operation (Rochette & Eriksen-Hamel 2008; Fang et al. 1998; Pumpanen et al. 2004; Rayment & Jarvis 1997), and methodological inter-comparison studies have attempted to

harmonize the outputs from disparate methods for collecting and analyzing gas emissions from the soil surface (Butnor et al. 2005; Pumpanen et al. 2003). Similarly, effort has been made at the theoretical level to describe the relationship between fluxes and environmental variables such as soil temperature, moisture and management, allow the interpolation and/or stratification of fluxes, and reducing the sample size needed for measurements accordingly (Rochette et al. 1991; Xu & Qi 2001; Lin et al. 2011). Whilst these difficulties are not yet fully resolved, a complimentary approach is to develop a more efficient sampling strategy.

In soil science generally there is a significant amount of statistical guidance on the design of field experiments and surveys (Cochran 2007; John 1998) and this has served us well in our analysis of the effects of manipulative interventions and soil inventories. In trying to quantify soil GHG emissions, however, we face the simple practical constraint of sample size. The limited number of chambers (or collars) that can be deployed, the amount of time required for a single measurement (especially for CH_4 or N_2O fluxes), the limited number of gas samples that can be collected and analyzed (in off-line closed systems) or the limited number of chambers that can be multiplexed together (in open systems) all act to limit the number of locations that can realistically be sampled within any given project situation.

In some soil systems, particularly agricultural ones, intensive management has the effect of reducing spatial heterogeneity to manageable levels, thereby reducing the number of measurements required to capture population variance accurately. This is not generally true and spatial heterogeneity combined with limited sample size presents considerable opportunity for bias to enter into our measurements such that even when attempts are made to stratify sampling according to known sources of variance, uncertainty estimates remain large (Raupach et al. 2005).

A large number of samples are required to maintain the accuracy of measurements because of the high spatial variability of the GHG fluxes (Ambus & Christensen 1994; Dai et al. 2012; Rayment & Jarvis 2000; Rodeghiero & Cescatti 2008). For a finite population, the number of samples needed for a given error in the population mean can be derived by:

$$n = \frac{n_0 N}{n_0 + N - 1}, n_0 = \frac{z^2 CV^2}{E^2} \quad (1)$$

Where N is the population size, z =1.96 (for 95% confidence), E (%) is half-length of the confidence interval as a fraction of the population mean and *CV* is the coefficient of variation of the population. In practice, a pilot study or an investigation of historical data is necessary to estimate the *CV* (or at least establish an upper limit).

Constrained by several limitations such as labor effort, time and budget, the sample size required by simple randomized design is usually too large to apply in practice. Stratified sampling by vegetation or soil types (Fiener et al. 2012; Panosso et al. 2009; Schelde et al. 2012; Kreba et al. 2013), or topography (Imer et al. 2013; Fang et al. 1998) is widely used to reduce overall variance by applying simple random sampling to each strata. These stratifying methods may become invalid when the spatial variability of the GHG fluxes is controlled (even partially) by an unknown driver, or dominated by factors such as soil temperature and moisture that vary at the finest scale, even within strata (Rochette *et al*., 1991; Stoyan *et al*., 2000; Allaire *et al*., 2012). For these reasons, chambers have limited ability to measure accurately fluxes at such small scales and applying simple random sampling to each stratum may introduce large errors and biases in the estimate of population mean and variance.

With the aim of reducing measurement errors and biases associated with limited resources, here we present a staged approach to sampling that retains the essential characteristics

of the population distribution within a small sample size. Related, but less effective approaches have been investigated previously (Folorunso & Rolston 1984; Rodeghiero & Cescatti 2008). Our method (see details below) expands the approach used in (Rodeghiero & Cescatti 2008), where a heterogeneous field was divided into sub-regions by a pre-sampling stage, which reduced the total variance of the whole region. There is an extent to which our method can be viewed as a mathematical stratification, leading to a completely general sampling method. Drawing on published datasets of soil GHG emissions across a range of spatial variability, we show that this sampling strategy reduces uncertainty in all cases compared with the simple random sampling, and particularly where sources of variance are large.

2 Methods and Data

Two sampling strategies were modeled using simulation: (1) simple random sampling (SRS); (2) a resampling or two-stage sampling (2SS). Staged sampling consisted of an initial survey where a relatively large number of samples were made. Two descriptive statistics (mean & variance) were calculated for this set of samples; the mean is of primary concern when quantifying total GHG flux and variance is the critical factor revealing spatial variability. A Monte Carlo method and a cost function were then used to select a sub-sample from the 1st-stage sample such that the descriptive statistics of the sub-sample were closest to those of the 1st-stage sample. Three datasets with low, medium and high variability or coefficient of variance (*CV*) were explored. Three analyses were made to compare 2SS with SRS and investigate the effects of sample size on the improvements: (1) the error distributions of the final-stage sample mean and variance; (2) the effects of the final sample size on the root-mean-square errors (RMSEs) of the sample mean and variance; (3) the relation between the RMSEs and the initial and final sample size.

2.1 Assumptions and sampling strategies

For calculation purposes, we assume that the GHG emissions for a given area are discretized to a finite population size of *N*. For simple random sampling (SRS), *n* final samples are directly drawn randomly from *N*. The two-stage sampling (2SS) developed here invokes an extra initial sample set of size *n*' between the population and the final samples (Fig.1). The systematic/artificial errors between two independent measurements are assumed small enough to be negligble compared to the errors caused by the sampling methods as shown by (Mathieu et al. 2006).

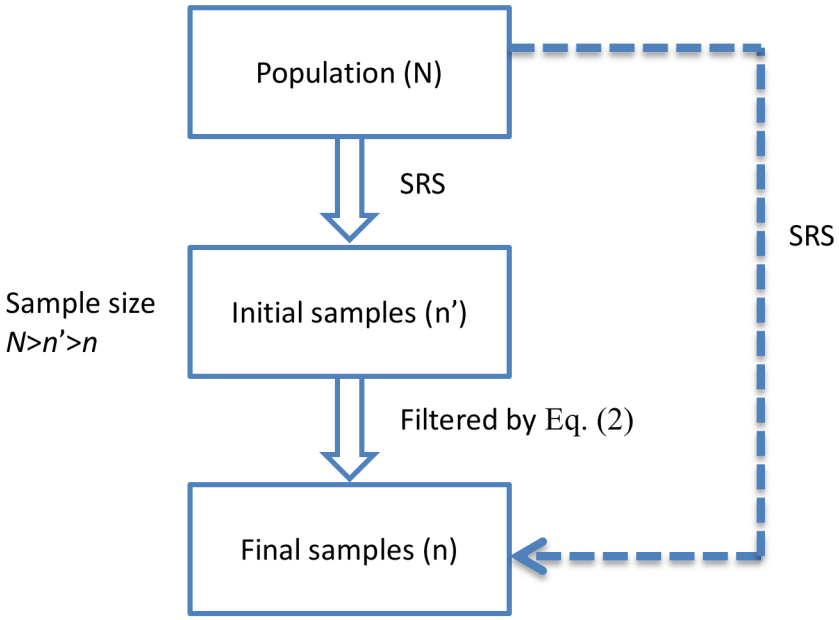

Figure 1 Workflow of the two-stage sampling (2SS), compared with the simple random sampling (SRS).

Note that $n' > n$; the filtering process from the initial sample to the final sample is performed by minimizing the cost or objective function:

$$min \quad f = \frac{|\mu_i - \mu'|}{\mu'} + \frac{|\sigma_i - \sigma'|}{\sigma'} - \frac{|\mu_i - \mu'|}{\mu'} \cdot \frac{|\sigma_i - \sigma'|}{\sigma'}, \quad i = 1,2,3, \dots, C_{n'}^{n} \qquad (2)$$

Where μ' and σ' are the mean and standard deviation of initial samples, while μ_i and σ_i are the mean and standard deviation for each combination of n from n' ($C_{n'}^{n}$). The aim of this process is to minimize f by finding the set of final samples that is most representative of the initial sample in terms of the errors of both mean and standard deviation. The choice of this cost function is pragmatic and may depend slightly on the subjective view of these statistics and purpose for which the data are collected. The function provided here selects a sample with a representative estimate of both the mean and standard deviation, the two most important sample features for atmospheric and biological modeling of GHG flux. Without suggesting that one is more important than the other, we assumed an equal weighting as implied in Eq. (2), however, in some cases a larger weight could be assigned to the sample mean if total flux is considered a higher priority. The first two terms in the function guarantee the choice of a sub-sample with a minimum sum of errors in mean and standard deviation and the third product term avoids choosing a sample with an extreme disparity between the first two terms, where one is overwhelmingly larger than the other. Weighting the mean and variance equally, this minimization process can be seen as a generalization of the method from a previous study (Rodeghiero & Cescatti 2008) where a simple approximation was achieved by stratifying the data by mean and variance and then selecting sub-samples at random from each strata.

In conventional two-stage sampling methods, the population is usually stratified into a sample of primary units from which a sample of secondary units is selected (Thompson 2012). While stratification is not explicit in 2SS, defining unequal strata and subsampling from these such that the sample represents the population would achieve similar results. The initial samples in 2SS can be considered as auxiliary information for improving the selection of the final

samples, similar to the way in which double or two-phase sampling adopts auxiliary information to improve inference of the population (Thompson 2012).

2.2 Datasets and statistics

The 2SS approach is illustrated using a dataset extracted from Mathieu *et al.* (2006) where 36 points were sampled at 3 m spacing on a 20 m × 20 m plot of a cultivated Gleyic luvisol located at Citeaux in the Saone river plain, near Dijon (Eastern France) in April 2003. We used these 36 samples as an adequate approximation of the population. The CO_2 flux dataset was used in this study and the flux unit was converted from g C ha^{-1} d^{-1} to g C m^{-2} d^{-1}.

In order to create datasets representing different degrees of variability without altering the mean, the original dataset A was expanded according to the following linear mapping,

$$y_i = (x_i - \mu) \cdot c + \mu, i = 1,2,3, \dots, 36 \qquad (3)$$

Where y_i are the new data points and x_i are the original ones. μ is the mean value of dataset A. The constant c ($\geqslant 0$) is an expanding factor. Two datasets with higher and lower *CV* were generated by setting $c = 2$ and $c = 0.5$ accordingly.

Note that the normality assumption for the distributions of the sample statistics is not appropriate for small sample size where the central limit theory becomes invalid. Therefore, here we simply employed the sample mean and variance as the estimators for the population mean and variance based on the method of moments (Feller 1968).

RMSE was used as an evaluator for the goodness of the sampling strategies. For example, the *RMSE* of the mean is defined by

$$RMSE\ of\ Mean = \sqrt{\frac{\sum_{i=1}^{m}(\bar{x}_i - \mu)^2}{m}} \qquad (4)$$

Where μ is the mean of the population and $\bar{x}_i$ are means of samples. m is the sampling repetition and was set to 1000 in our simulations to get a sampling distribution. *RMSE* is the square root of the mean squared error (*MSE*), which is a risk function corresponding to the expected value of the squared (quadratic) error loss and measures the estimator's bias.

RMSE of variance was calculated similarly, replacing the terms of the mean in Eq. (4) with the terms of the population and sample variance. In the remainder of this paper, the *RMSE* of mean and variance are designated as *RMSE*s for clarity.

3 Results

3.1 Error distributions of the sample mean and variance

We started with a simple case that a fixed initial sample size at 18 (n'=18) and a fixed final sample size at 6 (n=6) that are typically used per date. Distributions of the errors in the sample mean and variance were given in Fig. 2. Compared with SRS, application of 2SS resulted in general improvements in the accuracy of sample mean and variance for all datasets. A normally distributed error suggests a good sampling method and it was clear from the fitted normal curves (smooth lines in Fig. 2) that SRS error distributions were not normally distributed. This was confirmed by the Anderson-Darling (AD) test which showed that none of the error distributions from SRS should be accepted as normal (p<0.01) at the 5% significance level, highlighting the shortcomings of SRS in capturing the population's features when the sample size was small (e.g. 6 in this case). In fact, assuming normality for the distributions of sample statistics in a Monte Carlo estimate is not appropriate when the sample size is small, and can lead to a biased or erroneous inference to the whole population. On the contrary, when using 2SS, the errors of the sample mean did not fall into the critical regions for any of the datasets (p = 0.9376,

0.5202 and 0.7426), implying more unbiased and accurate estimates of the population mean. Distributions of the errors in the sample variance were non-normal for both methods, although clear improvements can be seen when 2SS was applied (Fig. 2d, 2e, 2f).

Bias and variance of a statistic estimator are used to quantify the amount of improvements in the sampling error and *RMSE* which incorporates both these aspects (i.e. *RMSE* can be written as the sum of the variance of the estimator and the bias of the estimator) is thus an appropriate evaluator. As shown in Table 1, the absolute reduction in *RMSEs* increased as the *CV* of the datasets increased, suggesting that the gain from 2SS may be proportional to the heterogeneity of the underlying population. In fact, the improvement in the *RMSE* of mean and variance were nearly proportional to the *CV* and CV^2 respectively (see below). Relative reductions in the two *RMSE*s were respectively around 55% and 58% for all datasets, demonstrating that the applying 2SS reduced the risk of getting unrepresentative samples by over 50% irrespective of the dataset's variability.

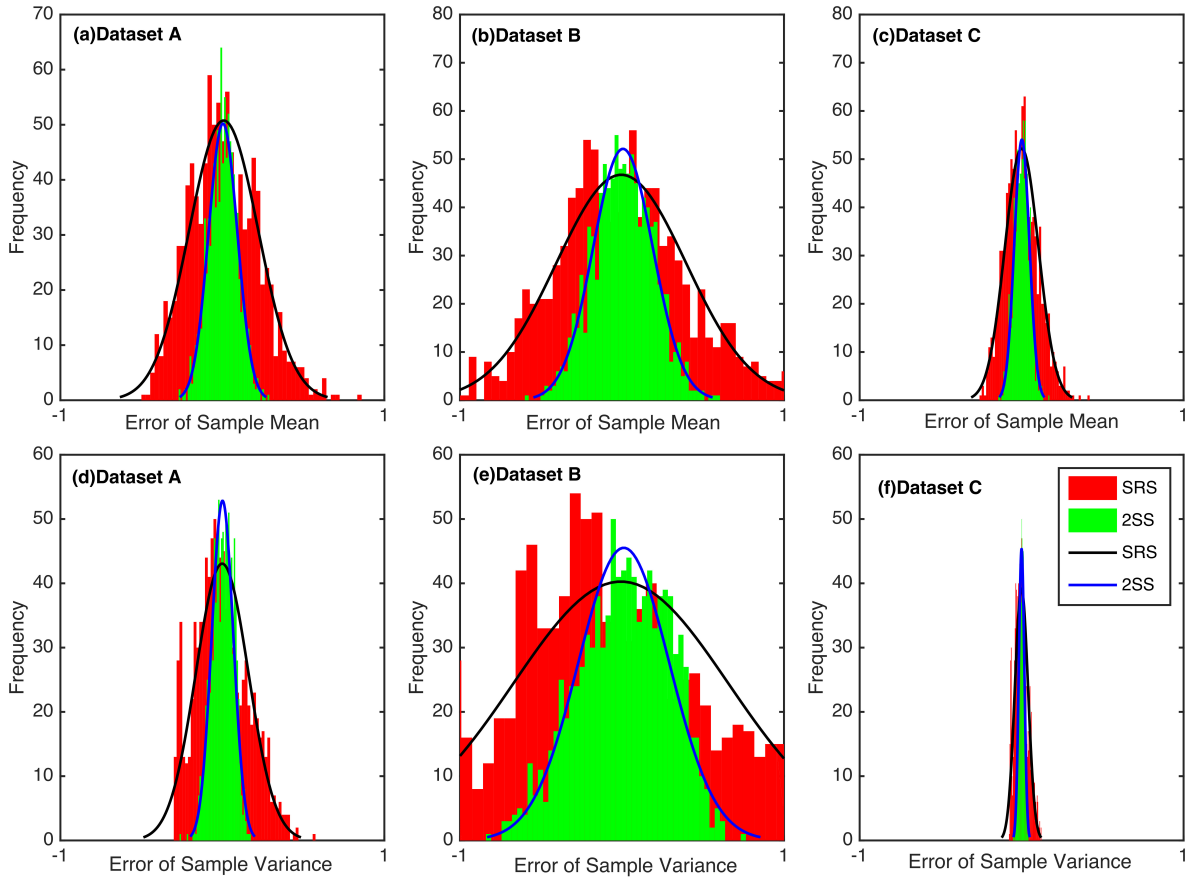

Figure 2. Error distributions of the sample mean and variance for the three datasets. The initial and final sample sizes were fixed at 18 and 6 respectively for 2SS. (a, b, c) Error distribution of the sample mean for dataset A, B and C. (d, e, f) Error distribution of the sample variance for dataset A, B and C.

Table 1. Improvements in RMSEs for the three datasets when using 2SS compared to SRS

Datasets	*CV*	*RMSE* (mean)		Absolute reduction	Relative reduction	*RMSE* (variance)		Absolute reduction	Relative reduction
		SRS	2SS			SRS	2SS		
A	0.603	0.202	0.090	0.112	55.6%	0.167	0.070	0.097	58.0%
B	1.205	0.420	0.199	0.221	52.5%	0.672	0.274	0.398	59.3%
C	0.301	0.102	0.046	0.056	54.6%	0.040	0.017	0.023	57.8%

3.2 Sample errors vs. the initial (n') and final (n) sample size

For a given initial sample size (*n'*), we can conduct a sensitivity analysis of the final sample size (*n*) to investigate how *RMSE*s vary with *n*. Without loss of generality, and to limit calculations to a manageable number in relation to the original dataset in Mathieu et al. (2006), *n'* was set to 18 while *n* ranged from 2 to 17. We calculated the *RMSE*s for the three datasets using the two sampling methods separately. As might be expected, datasets with higher *CV* produced larger *RMSE*s as shown in Fig. 3a&3b. *RMSE*s decreased gradually as *n* increased for SRS (dotted lines in Fig. 3a&3b) while for 2SS (solid lines with markers in Fig. 3a&3b), *RMSE*s remained almost constant for all final sample sizes greater than 2. This indicates that by using 2SS many fewer samples (e.g. 3) can achieve the same expected level of accuracy as a large sample number (i.e. 18 in this case) because of the efficacy from the combination of a larger initial sample size and the selection function Eq. (2). Additionally, in terms of the absolute difference between the two methods, dataset B (with the largest *CV*) showed the greatest decrease in *RMSE*s while dataset C showed the least, suggesting that the more heterogeneous sample area, the greater the absolute benefit of using 2SS. This result demonstrated that for 2SS, *RMSE*s were mainly determined by the initial samples and the cost function Eq. (2) performed well in selecting a set of samples with a reliable mean and variance.

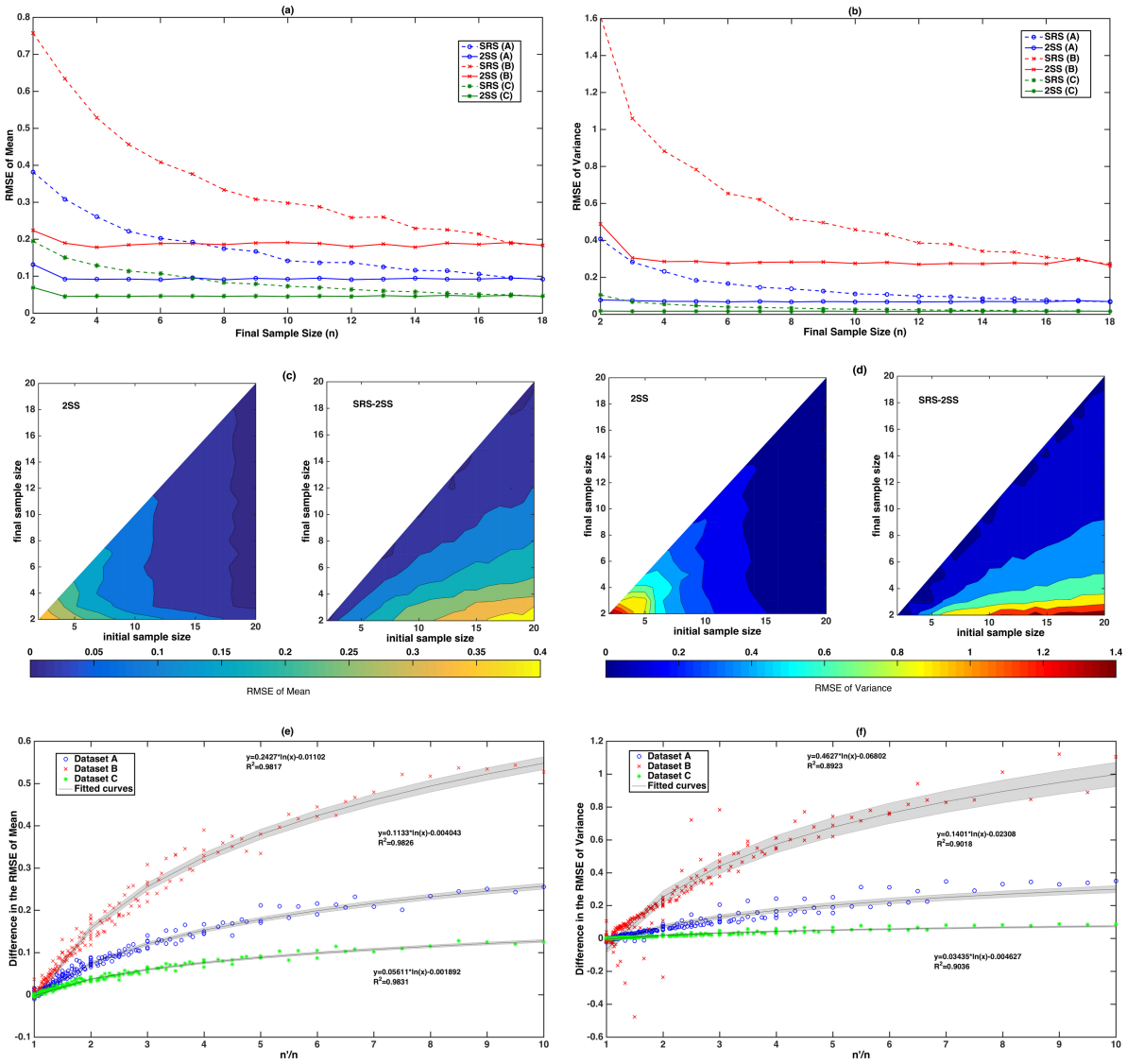

Figure 3. Sample errors vs. the initial (n') and final (n) sample size. (a, b) *RMSE* of mean (a) and variance (b) for the three datasets when $n' = 18$ and $2 \leqslant n \leqslant 17$. (c, d) The *RMSE* of mean (c) and variance (d) for all combinations of the initial and final sample sizes for dataset A. The difference in the *RMSE*s between SRS and 2SS (SRS-2SS) are shown on the right of each subplot. Only half of the graphic areas (triangle areas) are filled because n is necessarily $\leqslant n'$. (e, f) The difference in the RMSE of Mean (e) and Variance (f) against the ratio n'/n for the three

datasets. Logarithmic curves, y = a * ln (x) + b, were fitted and the grey shaded area represents 95% confidence bounds for the parameters a and b.

Recalculation of *RMSE*s was conducted for 2SS using every combination of n' and n satisfying $2 \leq n \leq n' \leq 20$ (resulting in 190 combinations in total) for each of the three datasets. Because the essential difference between the two methods is the filtering process represented by Eq. (2), SRS can thus be regarded as a quasi-two-stage sampling strategy without the filtering process, i.e. where the *RMSE*s will not vary with n'.

Unsurprisingly, except for the scale difference as indicated in Fig. 3a & 3b, the patterns of *RMSE*s were similar among the (related) datasets, therefore filled contours of *RMSE*s are only shown for dataset A in Fig. 3c & 3d. Compared to SRS, the *RMSE*s for 2SS were dominated by n' rather than n as the clear horizontal gradients show. Again, this suggests that the cost function, Eq. (2) worked well for selecting final samples that were representative of the initial ones. In other words, obtaining better accuracy in the estimate of mean and variance by applying a small n was achievable as long as n' and the cost function were chosen appropriately. 2SS worked to reduce *RMSE*s for nearly every combination of sample size as illustrated by the positive difference between SRS and 2SS (SRS-2SS) for all points on the plots. The average reduction in *RMSE*s was approximately 30% for all datasets.

The largest improvements were found at the lower right corner where the ratio, n'/n was large and appeared to decrease as the ratio n'/n decreased, suggesting a potential positive relation between them. A scatter plot of the difference in *RMSE* of the mean and n'/n showed a logarithmic relation ($y = a * \ln(x) + b$) for all datasets (Fig. 3e). The coefficient, a was found to be proportional to the dataset's *CV* while b was close to zero, suggesting the following,

$$\text{Difference in the } RMSE \text{ of the mean} = k * CV * \ln(n'/n) \qquad (5)$$

Where k is a constant number, e.g. 4.8 in this case. This provides us a good estimate of the gains obtainable by using 2SS and highlights that improvements increase proportionally with the variance of the population.

A similar relation was found between the difference in the *RMSE* of variance and the ratio n'/n (Fig. 3f). The coefficient a, however, was found to be proportional to the square of the dataset's CV, suggesting a similar function,

$$\text{Difference in the } RMSE \text{ of the variance} = k * CV^2 * \ln(n'/n) \qquad (6)$$

Here k is also a constant number, e.g. 3.2 in this case. Again, this result demonstrated that the improvement possible by employing 2SS rather than SRS can be quantified by basic functions, which can serve as guide to allocate limited measurement resources in practice.

4 Discussions and conclusions

Chamber-based measurement of GHG flux is straightforward and has many advantages compared to micrometeorological methods. In terms of capturing spatial variation, however, it is easy to make unreliable inferences about the spatial average by accidentally using spatially unrepresentative samples, particularly when the sample size is small. In this study we have demonstrated the effectiveness of a new sampling strategy, 2SS, in reducing sampling error in both the sample mean and spatial variance. We have further demonstrated that the expected benefits of this approach increase with increasing spatial variability.

By constructing an appropriate cost function (e.g. Eq. (2)), it was much easier to obtain a small set of final samples that was nonetheless representative of the initial samples, and provided an accurate estimate of the population mean and variance. Depending on the sample size chosen for the two stages, improvements in the sample mean and variance averaged 30% for all three

datasets used in this study. Compact relations were found between the potential benefits and the sample size ratio (n'/n), providing an easy guideline to allocate sampling resources (Fig. 3). Considering that the GHG flux datasets tend to be highly spatially heterogeneous as a result of diverse vegetation types, land-surface types and/or soil conditions (Reichstein 2003; Valentini 2003), SRS cannot be recommended (often) without a manageably large sample size. 2SS is a rather simple statistical method, leading to a more advanced sampling strategy that could contribute to improving GHG flux estimates irrespective of site or gas measured. In fact, the technique could be employed to reduce sample error in any situations where spatial variance is higher than a manageable sample number can effectively capture. The three datasets used in this study differed in their *CV* values, however, one should be aware that the *CV* alone as an index is not sufficient to completely characterize the heterogeneity of a source field, even though it has been the most widely-used and intuitive statistics (Buczko et al. 2015).

In the agricultural example used here, a history of management interventions such as ploughing and fertilization may have tended towards homogenizing the soil properties and microbial communities which can eventually "rectify" any areas of particularly high or low carbon-cycling activity. As such, this case (represented by dataset C with the lowest CV value) possibly represents the minimum benefit that 2SS confers compared to SRS. In more natural and unmanaged ecosystems, heterogeneity is typically higher and furthermore increases with time. This is most clearly seen in forest/woodland systems where the development of soil properties is highly influenced by proximity to individual trees even at fine scales (e.g. 12.5 cm, Jackson & Caldwell 1993) and in grass/sedge systems which over time become increasingly dominated by tussocks, particularly when the water-table is seasonally close to the surface (Soussana et al. 2007; Reynolds et al. 1997). Such cases represent a significant focus of GHG flux study and

would particularly benefit from the use of 2SS. Similarly, even in highly managed systems, such as livestock grazing systems, where the soil is initially highly uniform, large heterogeneity is found in the fluxes of GHGs (primarily N_2O and CH_4) associated with inputs from animal excretion (Saggar et al. 2004). Accurate quantification of these would also benefit significantly from the use of 2SS.

With the simple random sampling approach, the ideal sample size is always "as many as possible", but the "optimal" sample size varies with many factors as mentioned previously, e.g. the financial and human labor capacity, the study site, etc. To our knowledge, the complexity of making such resource allocation decisions has not been discussed previously, and the method presented here provides a statistical view on a fundamental issue that is often glossed over. We provide as concrete and robust a methodology to address the problem of sample size deficiency as it is possible to provide without consideration of the specific project requirements that the audience may encounter. The same applies to all statistical methods.

In 2SS, the increased effort spent in conducting a larger initial sampling at the beginning of the project becomes worthwhile in a long-term measurement through a significant reduction both in the likely sample error, and in the long-term effort required (e.g. fewer samples needed for accurate quantification). Nevertheless, this does raise a question about whether the second-stage sample remains the optimum sub-sample if the population changes with time. Fully understanding this requires new datasets with greater spatial and temporal resolution and is beyond the discussion of this paper, nevertheless, despite the assumption suggested in (Rodeghiero & Cescatti 2008) that GHG fluxes are temporally invariable over a specified period or maintain a relatively similar rate of change with time (e.g. high flux areas remaining high flux areas and vice versa), we suggest that a seasonal repetition of the initial survey should be

conducted to ensure that longer-term temporal variations are captured. For example, it is well known that the day-to-day variation in CO_2 flux is mainly driven by variations in light, temperature and water, but seasonal variation includes changes driven by, for instance, phenology. A few repetitions of the first stage sampling of 2SS at the critical stages of vegetation change (e.g. early and middle growing seasons) could update our choice of optimal sample location, thus increase the estimation confidence of flux.

Finally, here we have focused on two statistics, mean and variance, as these are the primary descriptors of a population. Nevertheless, these alone may not completely capture the true spatial pattern of GHG flux (i.e. two datasets with the same mean and variance may have different spatial patterns). This is particularly the case where the flux hotspots exist (Stoyan et al. 2000; Parkin 1987), especially for CH_4 and N_2O fluxes. Long-term spatial hotspots can increase the spatial heterogeneity significantly through microbial processes at microscale (less than 1m), such as denitrification (Farquharson & Baldock 2008; Groffman et al. 2009) and/or methanogenesis (Wachinger et al. 2000). SRS is likely to under sample events with low probability and thus is not recommended for capturing hotspots. With a large sample size at the initial stage, 2SS is more likely to catch rare events. The current form of 2SS can be further improved by including skewness or kurtosis in the cost function, which would allow the final sample to express similar population characteristics as the overall population. If this inclusion is found to be important, the most comprehensive spatial treatment would be to derive the variogram (Isaaks & Srivastava 1989) for the sample area, and select a sub-sample that expressed similar variogram parameters, i.e. consider the spatial autocorrelation of the datasets (Wang et al. 2012). This, however, would require a significantly larger and more detailed initial

samples to extract a reliable variogram function, and we leave this task to the next-stage research.

To conclude, SRS never outperforms 2SS, and 2SS always increases the sampling efficiency in the long term. Since it is a purely statistical model aimed at obtaining a better estimation of the population mean and variance, it can be easily applied to other datasets representing various types of land surfaces. Making the simplification that coefficient of variation (*CV*) is a reasonable measure of spatial heterogeneity, it is clear that the improvements gained through using 2SS are higher in more complex land surface types; the higher the *CV*, the higher the gain from using 2SS. Using a very simple form, the approach proposed here provides a statistical view on a very fundamental issue which should receive greater attention, and provides a concrete and robust methodology to address the problem of sample size deficiency.

Acknowledgments

This work was supported by China Scholarship Council - Bangor University PhD Scholarship Program. We thank anonymous peer reviewers and editors for their valuable comments and suggestions.

References

Allaire, S.E. et al., 2012. Multiscale spatial variability of CO2 emissions and correlations with physico-chemical soil properties. *Geoderma*, 170, pp.251–260. doi: 10.1016/j.geoderma.2011.11.019.

Ambus, P. & Christensen, S., 1994. Measurement of N2O emission from a fertilized grassland: An analysis of spatial variability. *Journal of Geophysical Research*, 99(D8), pp.16549–16555. doi: 10.1029/94JD00267.

Baldocchi, D.D., 2003. Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: Past, present and future. *Global Change Biology*, 9, pp.479–492. doi: 10.1046/j.1365-2486.2003.00629.x.

Baldocchi, D.D., Meyers, T.P. & Wilson, K.B., 2000. Correction of eddy-covariance measurements incorporating both advective effects and density fluxes. *Boundary-Layer Meteorology*, 97(3), pp.487–511. doi: 10.1023/A:1002786702909.

Buczko, U. et al., 2015. Spatial variability at different scales and sampling requirements for in situ soil CO_2 efflux measurements on an arable soil. *Catena*, 131, pp.46–55. doi: 10.1016/j.catena.2015.03.015.

Butnor, J.R., Johnsen, K.H. & Maier, C.A., 2005. Soil properties differently influence estimates of soil CO2 efflux from three chamber-based measurement systems. *Biogeochemistry*, 73(1), pp.283–301. doi: 10.1007/s10533-004-4022-1.

Cochran, W.G., 2007. *Sampling Techniques* Third Edit., New York.

Dai, Z. et al., 2012. Effect of assessment scale on spatial and temporal variations in CH_4, CO_2, and N_2O fluxes in a forested wetland. *Water, Air, & Soil Pollution*, 223(1), pp.253–265. doi: 10.1007/s11270-011-0855-0.

Davidson, E.A. et al., 2002. Minimizing artifacts and biases in chamber-based measurements of soil respiration. *Agricultural and Forest Meteorology*, 113(1), pp.21–37. doi: 10.1016/S0168-1923(02)00100-4.

Fang, C. et al., 1998. Soil CO_2 efflux and its spatial variation in a Florida slash pine plantation. *Plant and soil*, pp.135–146. doi: 10.1023/A:1004304309827.

Farquharson, R. & Baldock, J., 2008. Concepts in modelling N2O emissions from land use. *Plant and Soil*, 309(1-2), pp.147–167. doi: 10.1007/s11104-007-9485-0.

Feller, W., 1968. *An Introduction to Probability Theory and Its Applications*, doi: 10.2307/1266435.

Fiener, P. et al., 2012. Spatial variability of soil respiration in a small agricultural watershed — Are patterns of soil redistribution important? *Catena*, 94, pp.3–16. doi: 10.1016/j.catena.2011.05.014.

Foken, T. et al., 2004. Post-Field Data Quality Control. In *Handbook of Micrometeorology*. pp. 181–208. doi: 10.1007/1-4020-2265-4_9.

Folorunso, O.A. & Rolston, D.E., 1984. Spatial Variability of Field-Measured Denitrification Gas Fluxes. *Soil Science Society of America Journal*, 48, pp.1214–1219. doi: 10.2136/sssaj1984.03615995004800060002x.

Goulden, M.L. et al., 1996. Measurements of carbon sequestration by long-term eddy covariance: Methods and a critical evaluation of accuracy. *Global Change Biology*, 2(3), pp.169–182. doi: 10.1111/j.1365-2486.1996.tb00070.x.

Groffman, P.M. et al., 2009. Challenges to incorporating spatially and temporally explicit phenomena (hotspots and hot moments) in denitrification models. *Biogeochemistry*, 93(1-2), pp.49–77. doi: 10.1007/s10533-008-9277-5.

Gu, L. et al., 2012. The fundamental equation of eddy covariance and its application in flux measurements. *Agricultural and Forest Meteorology*, 152, pp.135–148. doi: 10.1016/j.agrformet.2011.09.014.

Hollinger, D.Y. & Richardson, A.D., 2005. Uncertainty in eddy covariance measurements and its application to physiological models. *Tree physiology*, 25(7), pp.873–885. doi: 10.1093/treephys/25.7.873.

Imer, D. et al., 2013. Temporal and spatial variations of soil CO_2, CH_4 and N_2O fluxes at three differently managed grasslands. *Biogeosciences*, 10(9), pp.5931–5945. doi: 10.5194/bg-10-5931-2013.

IPCC, 2000. *Land Use, Land-Use Change, and Forestry*, doi: 10.2277/0521800838.

Isaaks, E.H. & Srivastava, R.M., 1989. *An introduction to applied geostatistics*, New York: Oxford University Press.

Jackson, R.B. & Caldwell, M.M., 1993. Geostatistical Patterns of Soil Heterogeneity around Individual Perennial Plants. *Journal of Ecology*, 81(4), pp.683–692. doi: 10.2307/2261666.

John, P.W.M., 1998. *Statistical Design and Analysis of Experiments*, Philadelphia, PA: Society for Industrial and Applied Mathematics.

Jones, S.K. et al., 2011. Nitrous oxide emissions from managed grassland: a comparison of eddy covariance and static chamber measurements. *Atmospheric Measurement Techniques Discussions*, 4(1), pp.1079–1112. doi: 10.5194/amt-4-2179-2011.

Kreba, S. a. et al., 2013. Spatial and Temporal Patterns of Carbon Dioxide Flux in Crop and Grass Land-Use Systems. *Vadose Zone Journal*, 12(4). doi: 10.2136/vzj2013.01.0005.

Lin, X. et al., 2011. Response of ecosystem respiration to warming and grazing during the growing seasons in the alpine meadow on the Tibetan plateau. *Agricultural and Forest Meteorology*, 151(7), pp.792–802. doi: 10.1016/j.agrformet.2011.01.009.

Maljanen, M. et al., 2010. Greenhouse gas balances of managed peatlands in the Nordic countries - present knowledge and gaps. *Biogeosciences*, 7(9), pp.2711–2738. doi: 10.5194/bg-7-2711-2010.

Mathieu, O. et al., 2006. Emissions and spatial variability of N_2O, $N_2$and nitrous oxide mole fraction at the field scale, revealed with ^{15}N isotopic techniques. *Soil Biology and Biochemistry*, 38(5), pp.941–951. doi: doi:10.1016/j.soilbio.2005.08.010.

Oren, R.A.M. et al., 2006. Estimating the uncertainty in annual net ecosystem carbon exchange: Spatial variation in turbulent fluxes and sampling errors in eddy-covariance measurements. *Global Change Biology*, 12(5), pp.883–896. doi: 10.1111/j.1365-2486.2006.01131.x.

Panosso, a. R. et al., 2009. Spatial and temporal variability of soil CO2 emission in a sugarcane area under green and slash-and-burn managements. *Soil and Tillage Research*, 105(2), pp.275–282. doi: 10.1016/j.still.2009.09.008.

Parkin, T.B., 1987. Soil microsites as a source of denitrification variability. *Soil Science Society of America Journal*, 51, pp.1194–1199. doi: 10.2136/sssaj1987.03615995005100050019x.

Pumpanen, J. et al., 2004. Comparison of different chamber techniques for measuring soil CO_2 efflux. *Agricultural and Forest Meteorology*, 123(3), pp.159–176. doi: 10.1016/j.agrformet.2003.12.001.

Pumpanen, J. et al., 2003. Seasonal patterns of soil CO_2 efflux and soil air CO_2 concentration in a Scots pine forest: comparison of two chamber techniques. *Global Change Biology*, 9(3), pp.371–382. doi: 10.1046/j.1365-2486.2003.00588.x.

Raupach, M.R. et al., 2005. Model-data synthesis in terrestrial carbon observation: methods, data

requirements and data uncertainty specifications. *Global Change Biology*, 11(3), pp.378–397. doi: 10.1111/j.1365-2486.2005.00917.x.

Rayment, M.B. & Jarvis, P.G., 1997. An improved open chamber system for measuring soil CO_2 effluxes in the field. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 102(D24), pp.28779–28784. doi: 10.1029/97JD01103.

Rayment, M.B. & Jarvis, P.G., 2000. Temporal and spatial variation of soil CO2 efflux in a Canadian boreal forest. *Soil Biology and Biochemistry*, 32(1), pp.35–45. doi: 10.1016/S0038-0717(99)00110-8.

Reichstein, M., 2003. Modeling temporal and large-scale spatial variability of soil respiration from soil water availability, temperature and vegetation productivity indices. *Global Biogeochemical Cycles*, 17(4), p.1104. doi: 10.1029/2003GB002035.

Reth, S., Göckede, M. & Falge, E., 2005. CO_2 efflux from agricultural soils in Eastern Germany - comparison of a closed chamber system with eddy covariance measurements. *Theoretical and Applied Climatology*, 80(2-4), pp.105–120. doi: 10.1007/s00704-004-0094-z.

Reynolds, H.L. et al., 1997. Soil Heterogeneity and Plant Competition in Anannual Grassland. *Ecology*, 78(7), pp.2076–2090. doi: 10.1890/0012-9658(1997)078[2076:SHAPCI]2.0.CO;2.

Rochette, P., Desjardins, R. & Pattey, E., 1991. Spatial and temporal variability of soil respiration in agricultural fields. *Canadian Journal of Soil Science*, 196(90). doi: 10.4141/cjss91-018.

Rochette, P. & Eriksen-Hamel, N.S., 2008. Chamber measurements of soil nitrous oxide flux: are absolute values reliable? *Soil Science Society of America Journal*, 72(2), pp.331–342. doi: 10.2136/sssaj2007.0215.

Rodeghiero, M. & Cescatti, A., 2008. Spatial variability and optimal sampling strategy of soil respiration. *Forest Ecology and Management*, 255(1), pp.106–112. doi: 10.1016/j.foreco.2007.08.025.

Saggar, S. et al., 2004. A review of emissions of methane, ammonia, and nitrous oxide from animal excreta deposition and farm effluent application in grazed pastures. *New Zealand Journal of Agricultural Research*, 47(4), pp.513–544. doi: 10.1080/00288233.2004.9513618.

Schelde, K. et al., 2012. Spatial and temporal variability of nitrous oxide emissions in a mixed farming landscape of Denmark. *Biogeosciences*, 9(8), pp.2989–3002. doi: 10.5194/bg-9-2989-2012.

Soussana, J.F. et al., 2007. Full accounting of the greenhouse gas (CO2, N2O, CH4) budget of nine European grassland sites. *Agriculture, Ecosystems and Environment*, 121(1-2), pp.121–134. doi: 10.1016/j.agee.2006.12.022.

Stoyan, H. et al., 2000. Spatial heterogeneity of soil respiration and related properties at the plant scale. *Plant and Soil*, 222(1-2), pp.203–214. doi: 10.1023/A:1004757405147.

Thompson, S.K., 2012. *Sampling*, Hoboken: John Wiley & Sons.

Valentini, R., 2003. *Fluxes of carbon, water and energy of European forests*, Heidelberg:

Springer.

Venterea, R.T., Spokas, K.A. & Baker, J.M., 2009. Accuracy and precision analysis of chamber-based nitrous oxide gas flux estimates. *Soil Science Society of America Journal*, 73(4), pp.1087–1093. doi: 10.2136/sssaj2008.0307.

Wachinger, G., Fiedler, S. & Roth, K., 2000. Variability of soil methane production on the micro-scale: spatial. *Soil Biology & Biochemistry*, 32(8-9), p.1121. doi: doi:10.1016/S0038-0717(00)00024-9.

Wang, J.-F. et al., 2012. A review of spatial sampling. *Spatial Statistics*, 2, pp.1–14. doi: 10.1016/j.spasta.2012.08.001.

Xu, M. & Qi, Y., 2001. Soil-surface CO2 efflux and its spatial and temporal variations in a young ponderosa pine plantation in northern California. *Global Change Biology*, 7, pp.667–677. doi: 10.1046/j.1354-1013.2001.00435.x.